DOCUMENT RESUME

ED 152 844 95 TM 007 059

AUTHOR Godbout, Robert C.
TITLE The Problem of Chance Significant Findings in Educational Research. Research Report No. 1.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.; Office of Education (DHEW), Washington, D.C.
PUB DATE Aug 75
CONTRACT NE-C-00-3-0066; NE-C-74-0088; OE-6-10-108
NOTE 15p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Educational Experiments; Educational Research; Predictor Variables; Probability; *Reliability; *Research Design; *Research Problems; *Statistical Analysis; *Tests of Significance
IDENTIFIERS Aptitude Treatment Interaction; Type II Errors

ABSTRACT

Exploratory research with large numbers of variables and even larger numbers of relationships is likely to result in frequent Type II errors; that is falsely accepting an incorrect null hypotheses. There are three ways of reducing Type II errors: (1) choosing a lower significance level, which then increases the probability of Type I error; (2) replicating the experiment, which is usually expensive; and (3) applying second-order significance tests. The procedures for applying second-order significance tests are described and exemplified by analysis of data from an aptitude-treatment interaction experiment. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SCOPE OF INTEREST NOTICE

The ERIC Facility has assigned this document for processing to: TM SP

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

ED152844

IJTR&D report

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EVALUATION OF TEACHING PROJECT

TM007 059

THE RESEARCH AND DEVELOPMENT CENTER FOR TEACHER EDUCATION • THE UNIVERSITY OF TEXAS AT AUSTIN

The *Evaluation of Teaching Project*—one of four projects at the Research and Development Center for Teacher Education—has as its mission to develop materials and strategies for teacher training, research and evaluation. The goals of the *Evaluation of Teaching Project* are to develop (1) a conceptual framework for the evaluation of teaching, (2) a sourcebook of validated teacher evaluation and research instruments, and (3) strategies for the evaluation of teacher trainee programs. These goals are being carried out with funds from the National Institute of Education.

In the process of meeting its objectives, the *Evaluation of Teaching Project* conducts systematic research in teacher behavior for the purpose of validating instruments and identifying characteristics of effective teaching. The following report describes one facet of this research. A complete listing of studies in this report series is available by writing to the Evaluation of Teaching Project, R&D Center for Teacher Education, University of Texas at Austin, 78712.

The Problem of Chance Significant Findings in Educational Research

Robert C. Godbout

The University of Texas at Austin

Research Report No. 1

August, 1975

The research reported herein was supported in part by National Institute of Education Contracts NE-C-74-0088 and NE-C-00-3-0066 and U. S. Office of Education Contract OE-6-10-108, Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education of the Office of Education an no official endorsement by those offices should be inferred.

The Problem of Chance Significant Findings in Educational Research

Robert C. Godbout

Inferential statistical analyses provide (1) descriptions of the strength and form of an observed relationship and (2) an estimate of the chance probability for that relationship. The chance probability is the probability of obtaining the observed relationship when no true relationship exists. According to the classical Neyman-Pearson theory of hypothesis testing, a significance level (α) is chosen a priori. If the chance probability for a particular finding is less than α, then that finding is taken as significant or valid--i.e., the finding probably reflects a true relationship. On the other hand, if the chance probability for a finding is greater than α, then that finding is assumed to be nonsignificant or spurious.

Findings are categorized as either significant or nonsignificant. However, the distinction between these two categories is not true relationship (significance) vs. no relationship (nonsignificance). Rather, a significant finding indicates that there is a high probability ($1 - \alpha$) that a true relationship exists while a nonsignificant finding indicates a lower probability that a true relationship exists. An observed finding associated with a small chance probability (p) is not certain evidence for a true relationship. Such a finding occurs by chance with probability p. As long as p is not 0, there is a chance that the finding is spurious. Significant findings can be spurious or chance-produced--the probability of such an occurrence being α.

There are certain cases in which a significant finding can be recognized as chance-produced and discounted. Consider the case of an experiment which produces a significant result but, in nine subsequent replications, no significant

results are obtained. Certainly our confidence in the initial results is shaken. The obvious inference would be to attribute this initial significant finding to chance. In contrast, if several of the replications had produced significant findings similar to the initial one, then our confidence in the initial result would have been strengthened.

Now consider the case in which we have ten different criterion variables all measuring the same characteristic. Interest is focused on whether two groups of subjects differ with regard to each of these variables. If the difference between the two groups is significant with regard to one of the criterion variables but not the nine others, we would probably conclude that we are dealing with a chance effect. As the number of obtained significant differences increases, the hypothesis that the results are produced entirely by chance becomes more and more questionable.

All researchers who depend upon statistical analyses must live with the possibility of chance-produced significant results. The probability of encountering such results increases with the size of the research design. The laboratory experimenter who performs only a few statistical tests need have relatively little concern for chance-produced significant results. His confidence in obtained significant results is quite high, this confidence usually approaching $1 - \alpha$. In contrast, the educational researcher employing a large-scale exploratory design must come firmly to grips with the problem. The present author is familiar with a number of educational research studies which have performed several thousand statistical analyses on a data set. If 2,000 analyses are performed and a .05 significance level chosen, then an average of 100 analyses are expected to be significant on the basis of chance. Say that 300 of the 2,000 analyses are actually found to be significant (such a low "hit rate" is not unexpected in exploratory educational research). To

accept these 300 results without further ado is akin to not separating the dirt from the potatoes. On the average, the probability that a significant result chosen at random actually reflects a true relationship is about .67 and not .95. Confidence in the significant results has been eroded due to the existence of chance-produced significance. Clearly an attempt should be made to eliminate the chance-produced significant results.

That chance-produced significance can erode our confidence in significant results has long been recognized by mathematical statisticians (e.g., Fisher, 1935). Multiple comparison techniques for analysis of variance represent attempts to deal with the problem of chance-produced significance. The strategy of these techniques is to counter the erosion in confidence by lowering α below the standard level (e.g., .05) and thereby raising confidence in significant results. Lowering α represents a possible approach to the chance significance problem in educational research. Alternative approaches are replication and second-order significance testing. Each of these three approaches will be discussed in turn.

Lowering α. Employing a lowered α level (say .01 or .001 rather than the usual .05) will lessen the problem of chance significance. However, this method has a critical disadvantage. Lowering α causes statistical power (the ability to detect true relationships) to decrease. Since statistical power in educational research is low to begin with due to measurement error (Cohen, 1969), it is the present author's guess that any decrease in power is undesirable. (To pursue a previous analogy, this method may involve throwing out most of the potatoes with the dirt.) Thus, lowering α is not recommended.

Replication. Replication provides a particularly elegant solution to the problem of chance significance. If a study is replicated, then the valid relationships should be those which are reliably significant--i.e., which

prove to be significant in both replications. The chance probability that a finding will be significant at the .05 level in both of two replications is only .0025. If each replication involves 2,000 analyses, then only 5 are expected to be simultaneously significant by chance in both replications. Replication provides the best solution to the chance significance problem. However, this method does possess a practical drawback--the prohibitive cost of replicating a large-scale educational research study.

Second-order significance testing. This relatively cheap method of dealing with chance-produced significance is recommended for use in any large-scale educational research study. It is not intended as an alternative to the sounder method of replication but rather as a preliminary test to determine the feasibility of a replication. The logical basis for second-order significance testing rests on the assumption that the reliability (validity) of a significant result can be guessed at by using methods other than the test-retest (replication) method. Test construction specialists have devised methods of estimating reliability from internal consistency. A similar approach can be used to assess the reliability of significant results. An isolated significant result is probably chance-produced while a significant result which is part of a pattern of significant results probably reflects a true relationship.

The second-order significance testing method involves examination of families of analyses. A family of analyses is a collection of related analyses. Examples of families of analyses include: (1) all analyses in a particular study, (2) all analyses involving variables from a particular criterion instrument, (3) all analyses involving a particular criterion variable from that instrument, and (4) all analyses involving tests of curvilinear predictor terms with regard to that criterion variable. Note that these four example families differ with regard to generality (inclusiveness)--(1) being the most general and (4) being the least general.

Consider a single family containing N analyses. This family is examined in the following manner:

(1) The number of chance-produced significant results expected in the family is calculated by multiplying N times α (the significance level).

(2) The actual observed number of significant results is compared to this expected number.

(3) If the observed number significantly exceeds the expected number (second-order significance test), then all significant results within that family are retained.

(4) If the observed number does not significantly exceed the expected number, then all significant results within that family are discarded as chance produced.

The binomial distribution is used to determine the probability of the second-order significance test (comparison of observed number of significant results to expected number of significant results). Note that the significant analyses within a family are discarded as chance produced when this probability is non-significant (i.e., greater than α). Details for determination of the probability of the second-order significance test are given in Appendix A.

The application of second-order significance testing is most easily illustrated by an example. Consider an investigation involving 100 predictor variables (50 teacher characteristics and 50 pupil characteristics) and 20 criterion scores (10 different measures of math achievement and 10 different measures of reading achievement). The following families of analyses are considered: (1) the family consisting of the entire set of 2,000, (2) the family consisting of the 1,000 analyses relevant to math achievement, (3) the family consisting of the 1,000 analyses relevant to reading achievement, (4) the family consisting of the 1,000 analyses relevant to teacher characteristics, (5) the

family consisting of the 1,000 analyses relevant to pupil characteristics, (6) the 100 families relevant to the 100 predictor variables, and (7) the 20 families relevant to the 20 criterion variables. Each of these 125 families is examined with regard to second-order significance. After all examinations are completed, the significant results contained in any family that did not reach second-order significance are discarded as chance produced. The remaining significant findings are assumed to be valid--i.e., they are the significant findings most likely to be replicable.

One shortcut in these procedures is available. If Family A is a subset of Family B and Family B fails to achieve second-order significance, then it is unnecessary to examine Family A. In this case, all significant analyses contained in A will be discarded on the basis of the failure of B to achieve second-order significance. In the above example, if the family involving all analyses relevant to teacher characteristics does not achieve second-order significance, then it is not necessary to examine the 100 families relevant to the 100 different teacher characteristics. Note that to maintain the possibility of such shortcuts, the most general (inclusive) families should be examined first and the least general families should be examined last.

This ends the description of the second-order significance approach to the problem of chance significant results. While these procedures may seem complex, they merely represent a formalized approach to identifying patterns or trends in the observed significant results. The Evaluation of Teaching Project has recently completed data collection in a large-scale replicated study. The results of this study will be used to test the validity of the second-order significance procedures. The procedures will be applied to the first replication results. The results of the second replication will then be examined to determine if the

significant results retained as valid in the first replication tend to again be significant in the second replication.

Previous application of second-order significance testing. Borich, Godbout, Peck, Kash and Poynor (1974) used second-order significance testing on the results of a large-scale study of aptitude-treatment interactions (ATI's). An ATI occurs when the criterion on aptitude regressions for two treatments are significantly different. This investigation involved two different teacher training programs (treatments), 22 personality (aptitude) variables, and 111 classroom behavior and attitude measures (criteria). A total of 2,127 ATI analyses (all possible aptitude-criterion pairings were not employed) were performed with α chosen at .10. The number of expected significant results on the basis of chance was 212.7 which 269 were actually obtained. Second-order significance techniques were employed and 162 significant results were discarded as chance produced. Of the remaining 107 significant findings 96 involved aptitude and criterion variables which can easily be interpreted in terms of social desirability. The consistency of these 96 results was quite amazing. A total of 84 were such that the Treatment 1 criterion on aptitude regression (when both criterion and aptitude were scaled from low to high social desirability) was less positive than the Treatment 2 regression. Furthermore, in 74 cases the Treatment 1 regression slope was negative while the Treatment 2 regression slope was positive. The surprising consistency of these results tends to confirm the validity of second-order significance testing.

References

Borich, G. D., Godbout, R. C., Peck, R. F., Kash, M. M., & Poynor, L. H. Final report: An evaluation of the personalized model of teacher training. Research and Development Center for Teacher Education, The University of Texas at Austin, 1974.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Fisher, R. A. The design of experiments. London: Oliver I. Boyd, Ltd., 1935.

McNemar, Q. Psychological statistics. New York: Wiley, 1966.

Appendix A

Determination of the Probability for the Second-Order Significance Test (Comparison of Observed Number of Significant Results with Expected Number of Significant Results)

Note. The following exposition assumes that the N analysis within a family are independent. This is rarely the case. Dependencies will exist in a family of analyses such that the probability of significance on an analysis is greater if other analyses are significant rather than nonsignificant. Similarly, the probability of nonsignificance on an analysis is greater if other analyses are nonsignificant rather than significant. Such dependencies will cause the probability of the second-order significance test to be underestimated. In fact, the calculated probability is the lower bound of the true probability, this lower bound being approached as the analyses approach independence. The significant analyses within a family are discarded as chance produced when the probability of the second-order significance test is too large (not smaller than α). Thus, the underestimate of the second-order significance test probability means that the second-order significance test is conservative in the case of dependent analyses. This circumstance is not considered to be damaging, since it will only increase the difficulty of discarding significant analyses as chance produced.

Consider a family of N analyses. Assume that the null hypothesis is true--i.e., there is no true relationship. Any obtained relationship will be entirely due to sampling error--i.e., to chance. Each of the N analyses can be conceived of as an independent event with two possible outcomes. The first possible outcome is the lack of a significant relationship at the α level, and the probability for the occurrence of this outcome is $1 - \alpha$. The second possible outcome is a significant relationship, and the probability of this outcome is α. The binomial expansion is applicable to this situation, in which we have N independent events with fixed-probability binary outcomes.

Say that a number (r) of the N analyses prove to be significant. Then we can use the binomial expansion to determine the chance probability of obtaining exactly r significant results and N - r nonsignificant results. The probability (P_r) of exactly r significant results can be calculated as follows:

$$P_r = \frac{N!}{r!(N - r)!} p^r q^{(N - r)}$$

where p is the probability of a significant result (i.e., α), and q is the probability of a nonsignificant result (i.e., $1 - \alpha$). In similar fashion, we can determine the chance probabilities for $r + 1$, $r + 2$, $r + 3$, . . ., $r + (N - r - 2)$, $r + (N - r - 1)$, and N significant results. Adding together the chance probabilities for r, $r + 1$, $r + 2$, $r + 3$, . . ., $r + (N - r - 2)$, $r + (N - r - 1)$, and N significant results, we obtain the chance probability for at least r significant results. Now if this chance probability is large (say greater than α), then the entire set of results can easily be attributed to chance. However, if the chance probability of at least r significant results is small (say less than α), then it is more difficult to discount the obtained significant differences as being due to chance. In this latter case, we reject the chance basis for the significant results. Obtaining at least that many significant results on the basis of chance is a relatively rare occurrence.

When N (the number of analyses) is large, then a normal curve approximation to the binomial distribution can be employed for determining second-order significance. This approximation is given by McNemar (1966, p. 61) as:

$$z = \frac{\text{obtained number} - \text{expected number}}{(Npq)^{\frac{1}{2}}} \qquad (1)$$

where p and q are the probabilities associated with the two outcomes, and N is the number of events. For present purposes, p is the probability of significance (α), and q is the probability for nonsignificance ($1 - \alpha$). The expected number is always pN. When r significant results are obtained, then a corresponding z value (normal deviate) can be calculated. The probability of at least r significant results is then approximated by the probability of obtaining a z at least as large--this latter probability being obtainable from a normal curve table.

McNemar's suggestions with regard to the application of this normal curve approximation are as follows. First, equation (1) is applied only when the expected number of significant results is greater than 10. Second, when the expected number of significant results is between 5 and 10, a correction for continuity is incorporated into equation (1). This correction for continuity involves subtracting .5 from the numerator in equation (1). Third, when the expected number of significant results is below 5, the normal curve approximation is not used.

When pN is less than 5, the chance probability for the number of significant results can be directly calculated using the binomial expansion. However, this tack is rejected for the present purposes. To understand why, consider what occurs when N is very small. If N is equal to 2 and α equal to .05, then the chance probability for obtaining two significant results is .0025. That is, when N equals 2, we must find two significant results in order to achieve second-order significance. Second-order significance is not achieved when only one result is significant--even though this represents one-half of the results.

Since the binomial expansion does not provide satisfactory results when pN is small (less than 5), a less efficient, simpler method of identifying chance results is employed in this case. This method consists of a simple comparison of the obtained number of significant results (r) with the expected number of significant results (pN). When r exceeds pN, then all significant results are accepted as valid. When r is equal to or less than pN, then all significant results are identified as chance results.